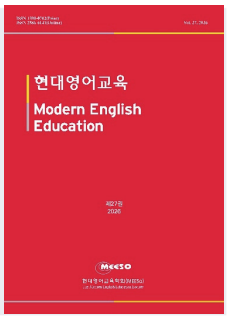
Modern English Education, Vol. 27, (2026) pp. 253-270
doi: https://doi.org/10.18095/meeso.2026.27.1.253

Modern English Education

ISSN: 1598-0782 / e-ISSN: 2586-6141
http://journal.meeso.or.kr



The Effect of Shadowing Activities Based on Dramatization of Picture Books on Chinese Young English Learners' English Proficiency

Yixin Xue (Korea University)
Chu Tang (Sichuan University)
Yanhua Wang (Chifeng Songshan No.6 Primary School)
Doseon Eur (Korea University)

Received: 6 March 2026
Revised: 20 March 2026
Accepted: 8 May 2026

Xue, Yixin, Tang, Chu, Wang, Yanhua, & Eur, Doseon. (2026). The effect of shadowing activities based on dramatization of picture books on Chinese young English learners' English proficiency. *Modern English Education*, *27*, 253-270.

Keywords
Chinese young English language learners, shadowing activities, picture books, English proficiency, English listening
중국 초등 영어 학습자, 섀도잉 활동, 그림책, 영어 능숙도, 영어 듣기 능력

Yixin Xue (First author)
PhD Candidate
Department of English Language Education
Korea University
2021011063@korea.ac.kr

Chu Tang (Co-author)
Research Assistant
Department of Emergency Medicine and West China Biomedical Big Data Center
Sichuan University
chu.tang@wchscu.cn

Yanhua Wang (Co-author)
English Instructor
Chifeng Songshan No.6 Primary School
Wangyanhua6@163.com

Doseon Eur (Corresponding Author)
Professor
Department of English Language Education
Korea University
doseon@korea.ac.kr

Abstract
This study investigates how shadowing activities based on the dramatization of English picture books affect the English proficiency of Chinese young learners. Fifty-eight fifth graders from a public school participated in a 10-week intervention, which included pre-, post-, and final- test measure. The experimental group utilized short audio recordings collected through a WeChat study group to verify task completion. Results showed that the experimental group achieved significantly higher scores on curriculum-based English assessments, as indicated by their Tongkao total scores. However, improvements in listening comprehension and long-term retention were limited. Notably, shadowing activities were paused for about a month between the post-test and final-test due to scheduling constraints at school, which may have contributed to the decrease in observed gains. These findings suggest that while combining dramatization with shadowing can lead to short-term improvements in academic English achievement in exam-oriented environments, continued engagement is essential for maintaining progress over time. The study highlights both the potential benefits and practical challenges of implementing supplementary oral practice for young EFL learners in under-resourced educational settings.

INTRODUCTION

English serves as a principal medium for international exchange across education, commerce, and culture, yet young learners in mainland China develop communicative competence unevenly. In many public primary schools, formal English instruction begins relatively late and is limited to a small number of forty-five-minute periods per week. These limited contact hours reduce opportunities for sustained input and purposeful oral practice, which are both known to constrain early language development. Beyond school, exam preparation often further narrows the scope of instruction, tending to emphasize decontextualized drills over sustained listening and speaking activities. Consequently, identifying scalable ways to extend input and meaningful practice—without increasing classroom contact time—has become a practical concern for educators and administrators alike (Butler, 2015).

The insufficient English class hours and exam-oriented teaching and learning settings constitute persistent instructional constraints. Importantly, the present study does not position exam-oriented assessment as a pedagogical target to be rejected or replaced. Rather, it treats exam-oriented conditions as an unavoidable institutional reality and examines whether communicative-oriented instructional practices can produce measurable benefits within existing assessment frameworks. Problems such as insufficient input, limited interactive practice, and instructional patterns shaped by test demands are common (Butler, 2015; Jin & Cortazzi, 2018). Under such instructional conditions, regular language hours at school are often used for repeated exam-oriented practice, which may constrain teachers' ability to sustain interactive tasks. It should be emphasized that the present study does not assume that exam-oriented structures can be readily removed from primary English education. Instead, these conditions are viewed as contextual limitations that emphasize the necessity of instructional strategies that can provide high-quality, contextualized, and replicable communicative input within the constraints of time (Feng, 2012; He, 2011; Hu, 2005; Ozverir et al., 2016).

High-stakes assessment intensifies these pressures. Empirical studies on washback in the Chinese context suggest that large-scale examinations are associated with increased attention to test-related practices, with mixed or uneven effects on classroom communication (Dong et al., 2023; Qi, 2005). Investigations of the National Matriculation English Test and related exam systems documents persistent tensions between communicative goals and examination requirements, which may limit opportunities for oral language development in everyday classroom practice (Cheng, 2008; Qi, 2007).

Given these constraints, dramatization may be a feasible option because it can produce rich, contextualized oral output within limited class time. According to Stinson and Winston (2011), dramatization activities provide students with opportunities to use and repeat high-frequency lexical chunks within meaningful narrative contexts. Picture book-based dramatization can support comprehension, chunk processing, and sensitivity to intonation and prosody, while enabling learners to produce more authentic oral output than traditional classroom activities (Chou, 2013; Serrurier-Zucker & Gobbé-Mévellec, 2014; Winston & Stinson, 2014; Wu, 2014). Because test preparation shapes the allocation of instructional time, dramatization represents a practical means of compressing contextualized communication into exam-constrained classroom settings. When children rehearse roles, deliver lines, and co-construct scenes, they encounter dense, repeated language anchored in story and emotion while practicing voice projection, intonation, and stance for an audience. However, dramatization in primary school typically depends on teacher orchestration and peer interaction; once class ends, its affordances are difficult to reproduce as solitary homework. This creates a gap between highly engaging in-class experiences and the routine, lower-quality practice that often follows after school (Stinson & Winston, 2011).

Therefore, dramatization could be viewed as a feasible instructional option in this context, as it supports communicative language use that may otherwise be difficult to sustain within traditional, exam-oriented classroom arrangements. However, because dramatization relies heavily on classroom interaction and teacher guidance, it requires supplementary forms of after-class practice, as such activities cannot easily be extended into individual after-school learning time. Accordingly, shadowing offers a complementary pathway for bridging this gap. In shadowing, learners repeat a spoken model immediately (or with minimal delay), thereby increasing time-on-task with comprehensible input while training prosodic accuracy (rhythm, stress, and intonation) and auditory discrimination. A growing body of research indicates that shadowing can enhance listening comprehension and refine perception-production mappings, with evidence particularly strong in secondary and tertiary contexts (Gill et al., 2014; Hamada, 2020; Irby et al., 2016; Parker et al., 2006). Its logistical advantages of audio-based delivery, brief sessions, and ease of monitoring make it feasible at scale as individualized homework (Hamada, 2015, 2018). In parallel, meta-analytic work on L2 pronunciation underscores the centrality of prosodic features for comprehensibility, suggesting that brief, frequent, prosody-focused practice may yield perceptible gains (Saito & Plonsky, 2019).

Despite separate lines of evidence supporting drama-based pedagogy in socially interactive classroom contexts and shadowing practice in individual listening-speaking training, the sequencing of these two pedagogical phases within a single

instructional design remains underexplored in primary EFL settings. Most prior studies treat dramatization and shadowing as independent instructional approaches rather than as functionally linked stages within a broader learning process.

Reviews of English education among young learners in East Asia repeatedly call for designs that extend input without intensifying teacher workload, especially in systems constrained by limited contact hours and strong assessment cultures (Butler, 2015). By testing an integrated model anchored in picture book narratives and operationalized through brief, audio-guided homework, the present study addresses this gap with an approach that is both instructionally coherent and administratively feasible.

In this quasi-experimental class-based study, both groups received classroom instruction centered on the dramatization of picture books that are in line with textbook themes. The experimental group additionally completed structured shadowing assignments after each lesson, using audio derived from the week's dramatized scenes and designed to highlight prosodic features and high-utility expressions. The integration aims to (a) prolong exposure to meaningful input, (b) strengthen perception-production links for prosody and formulaic language, and (c) examine whether such gains are reflected in measurable outcomes under existing assessment conditions. Grounded in this rationale, the study addresses the following questions:

RQ1. Do Chinese primary students in the dramatization plus shadowing condition show greater gains in English achievement (total English test scores) from pre-test to post-test than those in the dramatization-only condition?
RQ2. Do Chinese primary students in the dramatization plus shadowing condition show greater gains in listening comprehension from pre-test to post-test than those in the dramatization-only condition?

By articulating a mechanism that connects narrative-rich, in-class tasks to focused, prosody-oriented homework, the study contributes to ongoing efforts to design scalable, evidence-informed pedagogy for young EFL learners in China. (It also responds to calls for "convergence" in language education by bringing together literature-based instruction, task-based interaction, and speech-processing practice within a single, implementable model.)

THEORETICAL BACKGROUND

In order to conceptualize the relationship between dramatization and shadowing within a coherent instructional framework, the present study adopts a layered theoretical perspective in which classroom dramatization and post-class shadowing are understood as functionally distinct components within a broader learning sequence. Classroom dramatization aligns with sociocultural perspectives by providing opportunities for socially mediated language use, collaborative dialogue, and scaffolded participation in meaning-rich narrative contexts. Through role enactment and interaction, learners engage in externally supported performance that foregrounds discourse-level meaning, stance, and communicative intent.

Shadowing, however, is not positioned in this study as a sociocultural interactional activity in itself. Rather, it is conceptualized as an individual consolidation phase that follows socially mediated classroom experience. From a sociocultural standpoint, internalization involves the gradual transformation of socially mediated activity into individual control over linguistic forms. In this design, shadowing functions as a structured rehearsal mechanism that may support stabilization of forms, prosodic patterns, and auditory-motor coordination associated with previously dramatized language, rather than as a substitute for collaborative meaning negotiation.

Within this framework, dramatization and shadowing are treated as complementary phases that emphasize different dimensions of language development. Dramatization foregrounds socially situated meaning making, while shadowing emphasizes individual-level proceduralization and fluency-related consolidation. The integration of the two is therefore discussed in terms of sequencing and functional complementarity rather than theoretical equivalence. This framework provides the conceptual basis for examining their combined instructional effects in the present study.

LITERATURE REVIEW

This review synthesizes research in three areas: (1) the pedagogical benefits of dramatization in English as a Foreign Language (EFL) settings for young learners; (2) the impact of shadowing as a language learning strategy and post-class reinforcement activity; and (3) the integrated potential of combining dramatization and shadowing for more effective English instruction. Research in EFL contexts suggests that language learning is more effective when embedded in authentic

communicative settings, allowing learners to apply language meaningfully through direct engagement (Lave & Wenger, 1991). In language education, dramatization provides precisely such contexts by transforming reading materials into interactive performance tasks. Research has shown that drama enhances learner motivation, promotes deeper comprehension, and creates memorable, multimodal experiences (Ulas, 2008). By engaging visual, auditory, and kinesthetic channels simultaneously, dramatization fosters long-term retention of language forms. It also facilitates collaboration, critical thinking, and social interaction by situating language use in authentic communicative scenarios (Stinson & Winston, 2011).

Ma and Liu (2022) claimed that drama has been shown to reduce speaking anxiety, build confidence, and increase learners' willingness to communicate. Stinson and Freebody (2006) found that drama activities encourage risk-taking and help learners overcome the fear of making mistakes, an effect particularly salient in oral communication. In the Chinese context, however, the use of drama remains constrained by limited class time, curricular pressures, and a strong emphasis on exams. When instructional time is limited, instructors should consider how to maximize meaningful time-on-task for language exposure in classroom settings (Nation, 2007). Angelinawati (2019) demonstrated that dramatization is often treated as a supplementary activity rather than a core instructional practice, partly due to curriculum requirements and teachers' short classroom time, which reduces its potential impact on measurable language outcomes. Given the constraints of limited classroom time, dramatization alone may not be sufficient to sustain the repeated and systematic language practice required for measurable learning outcomes, thereby suggesting the potential value of additional practice beyond regular class instruction.

To address these limitations, shadowing has developed as a technique that reinforces auditory repetition while simultaneously engaging both receptive and productive skills. Shadowing, which originated in interpreter training, involves learners repeating spoken input with minimal delay, thus connecting perception, memory, and articulation in real time (Lambert, 1992). Cognitive psychology explains its benefits through Baddeley's model of working memory: the phonological loop temporarily stores auditory information, enabling learners to encode sounds and transform them into speech output (Baddeley, 2006). Because the shadowing activities were based on the previously enacted content, the cognitive load for the young language learners was reduced, making the activities manageable. Research on cognitive development suggests that basic working memory capacity develops relatively early in childhood (Diamond, 1995, 2013). Research has consistently demonstrated shadowing's potential in developing listening comprehension, prosodic awareness, and fluency. Hamada's large-scale experiments with Japanese EFL learners, for example, confirmed that shadowing promotes listening gains and improves learners' prosodic accuracy (Hamada, 2014). Meta-analyses of pronunciation instruction also highlight the central role of prosody in comprehensibility, suggesting that shadowing may be particularly effective when it targets rhythm and intonation (Saito, 2013; Saito & Plonsky, 2019; Saito, 2021). Furthermore, Swain's output hypothesis provides a theoretical rationale: by forcing learners to articulate what they hear, shadowing stimulates noticing, hypothesis testing, and metalinguistic reflection, processes essential for internalizing linguistic forms (Swain, 1993; Swain & Watanabe, 2019). Most existing studies on shadowing focus on secondary and tertiary learners, while developmental differences are substantial. Alloway, Gathercole, and Pickering (2006) claimed that young children demonstrate strong short-term retention capacity and may rely more on executive resources even for short-term memory tasks.

Nevertheless, several studies caution that shadowing has inherent limitations. Improvements are often short-lived without sufficient repetition and spacing, and cognitive fatigue may reduce learners' persistence. Takeuchi et al. (2021) found that classroom-based shadowing alone may not provide sufficient duration and intensity to support long-term retention, especially in prosody and vocabulary. Motivation is also a concern. Young learners, in particular, may perceive shadowing as mechanical and monotonous when it lacks contextual grounding (Uchihara et al., 2019). Without narrative engagement, shadowing risks becoming a decontextualized exercise, undermining its pedagogical effectiveness.

Within exam-oriented educational contexts, the feasibility of classroom practices becomes a central concern. The present study does not seek to challenge the legitimacy of exam response assessment systems; rather, it adopts a pragmatic stance by exploring instructional designs that can operate productively within such constraints. Dramatization and shadowing both rely on repeated exposure to language input, but their roles are distinct in language development. The former supports learners' discourse-level abilities through situational narratives and physical participation (Liu, 2002), whereas shadowing promotes real-time coordination between perception and production, particularly in relation to phonetic accuracy, intonation, and rhythm (Foote & McDonough, 2017). These differences reflect two different teaching approaches that emphasize various linguistic capabilities, rather than simply demonstrating the same underlying construct. At the same time, it is important to note that if people only focus on one method in isolation, in either way, limitations may emerge. To be more specific, each approach cannot maintain a balance between meaningful production and pronunciation at the same time. Dramatized instruction primarily stresses the meaningful reenactment of scenarios but may not systematically focus on precise word pronunciation. In contrast, shadowing could efficiently enhance phonetic skills in a relatively short time, yet it

fails to provide learners with meaningful contextual support, often leading to mechanical repetition over time. The present study, therefore, integrates both methods in an attempt to balance contextualized meaning-making and prosodic refinement within realistic instructional constraints.

Recent scholarship has begun to explore how to embed shadowing within richer instructional contexts. Dramatization, especially when based on picture books, provides such a foundation. Narrative contexts emotionally engage learners and supply repeated, high-frequency language forms. When shadowing is applied to the same dialogues or expressions performed in class, it reinforces both prosodic and semantic dimensions of language. Pereira et al. (2019) argue that emotionally meaningful input strengthens retention and motivates sustained engagement. By shadowing lines previously enacted in dramatization, young learners revisit familiar voices and characters, making the activity more personally relevant and less mechanical. Such practices may reduce processing demands. Moreover, this integration creates a pedagogical feedback loop. Dramatization introduces learners to authentic communicative situations and stimulates affective investment; shadowing consolidates these experiences through structured, individual practice beyond class. Together, the two methods are consistent with usage-based learning principles, which stress the importance of repeated, meaningful exposure for language development (Ellis, 2009).

Dramatization and shadowing each contribute uniquely to EFL learning but also display distinct limitations when used in isolation. Dramatization provides affective involvement and contextual immersion but lacks the systematic repetition required for automatization. Shadowing offers intensive practice but often suffers from motivational challenges without a meaningful narrative frame. Integrating the two creates a convergent model in which dramatization sparks engagement and provides authentic input, while shadowing ensures repeated rehearsal and consolidation. This synergy has the potential to extend learning beyond classroom walls, bridging the gap between contextual richness and procedural fluency, and thereby fostering durable improvements in young learners' English achievement.

METHODOLOGY

This study was conducted in a fourth-tier city in Inner Mongolia, China. Compared to megacities such as Beijing or Shanghai, Chifeng exhibits distinct disparities in public services, including education and healthcare. Yang (2006) notes that students in such regions often lack access to diverse learning resources, particularly in English language education. Opportunities for authentic language exposure, such as outdoor activities or culturally immersive events, are limited, leaving formal education and digitally mediated content as the primary means of English input. This narrowing of language exposure environments significantly reduces students' chances of using English in meaningful ways. Such contextual constraints may influence learners' engagement with technology-mediated learning practices and their responsiveness to after-class learning. At the same time, it is vital to establish conditions under which the intervention effects may be generalized beyond the immediate setting. Extending language practice activities beyond normal school hours becomes a critical methodological necessity due to limited regular English class hours, which constrain opportunities for authentic language use. Therefore, a face-to-face class-based quasi-experimental study was carried out during the fall semester of 2024 as part of the national "After-School Child Care Service" initiative.

Participants

The study involved 58 fifth-grade students (23 female, 35 male) from the same public primary school. All participants had been learning English since Grade 3 and had no history of overseas study or travel, suggesting broadly similar prior exposure to English. To reduce instructional variability, all students were taught by the same English teacher from Grade 3 throughout the research period. Participants were assigned to an experimental group and a control group (29 students per group) within the same intact class. The assignment was conducted at the student level within that class. No class-level random assignment, population-level random sampling, or formal stratified randomization procedure was implemented. Both groups received identical in-class instruction in terms of materials, pacing, and instructional procedures. The only systematic difference between groups was that the experimental group was assigned additional shadowing activities as part of after-class learning, while the control group did not receive this supplementary component.

The shadowing activities were implemented within the school's after-school program as a supplementary extension of classroom instruction. Details of the instructional procedures are provided in the Procedure section.

The study was conducted in accordance with institutional and school guidelines. Permission to conduct the study was obtained from the school administration. Informed consent was obtained from the participants and their parents or guardians

prior to data collection. All data were anonymized to ensure participants' confidentiality.

Instruments

In exam-oriented educational settings, instructional effectiveness is typically evaluated through standardized achievement measures. Therefore, in the present study, "English proficiency" is operationalized as curriculum-based English achievement, as reflected in performance on the district-level Tongkao assessment. Given the listening-focused nature of the intervention, changes in total English scores are interpreted as indirect indicators of curriculum-based academic performance rather than indicators of development across all language domains.

In China, high-stakes assessments such as the Gaokao (college entrance exam) and Zhongkao (high school placement exam) shape educational priorities. While primary school students are not subject to national-level standardized exams, many local education bureaus administer district-level assessments known as Tongkao. A paper-based Tongkao test was used to measure students' English achievement before and after the intervention. The test, administered annually in July to students in grades 3–6, includes a listening section (40 points) and a written section (60 points), for a total of 100 points. In the present study, overall English achievement is measured using the total Tongkao English test score, comprising both the written (60 points) and listening (40 points) components. The written portion comprises multiple-choice questions, fill-in-the-blank items, matching tasks, reading comprehension, true/false statements, and short essay prompts. All test items were drawn from the official item bank used by Chifeng's Education Bureau, with input from local English teachers to ensure age-appropriateness and alignment with the curriculum. In addition, the forms of pre-, post-, and final- test forms were the same.

To improve measurement transparency, Table 1 summarizes the internal consistency reliability of Tongkao subscales in the present sample. Tongkao consists of multiple parallel test forms with the same structure and scoring scheme, which are designed to have comparable difficulty levels across administrations. In this study, one representative form was selected for reliability estimation. As shown in Table 1, the overall objective section demonstrated acceptable internal consistency (α = .802), thereby supporting its use as a reliable indicator of curriculum-based English achievement within this sample. Subscale reliabilities varied, with listening and reading demonstrating moderate reliability and smaller subtests (e.g., pronunciation, grammar) showing lower alpha values, likely due to limited item numbers and heterogeneous item formats typical of school-based standardized assessments. Therefore, the relatively low alpha values for some subscales should be interpreted cautiously, as they likely reflect the brevity and mixed-format nature of these subtests rather than substantial problems with construct measurement. Reliability was not applicable to the writing task because it consisted of a single item.

In the present study, the Tongkao total score was used as the primary outcome measure to reflect curriculum-based English achievement under exam-oriented assessment conditions. The listening comprehension scores were analyzed as a secondary outcome to examine intervention effects more directly in response to the listening-focused nature of shadowing. The use of the total score does not imply uniform development across all subskills; rather, it reflects overall curriculum-aligned academic performance within the existing assessment framework. Although Tongkao is not specifically designed to assess prosodic processing, its listening section evaluates sentence-level and short discourse comprehension, which partially overlap with the auditory processing skills targeted in the structured shadowing activities.

TABLE 1

Consistency Reliability of Each Tongkao Subscale (Cronbach's α)

No.	Subscale	Item numbers	Total score	Total items	Cronbach's α
1	Listening	1-20, Part V	40	21	.603
2	Pronunciation	31–35	5	5	.399
3	Vocabulary	36–45	10	10	.688
4	Grammar	46–50	5	5	.331
5	Question Answer Matching	51–55	5	5	.862
6	Dialogue Completion	56–60	5	5	.856
7	Reading Comprehension	61–75	25	15	.624
8	Writing	76	5	1	Not applicable
9	Objective Section (Total)	1–20, Part V, 31–75	95	66	.802

Teaching Materials

The instructional materials used in this study were Level 3 graded readers from the Lisson Polaris series. These picture books were selected to supplement the standard English curriculum and to provide engaging, thematic content for dramatization and shadowing practice. In addition to thematic relevance, the selected books were considered linguistically appropriate for Grade 5 learners who had received approximately three years of formal English instruction. The texts predominantly feature high-frequency vocabulary, short clause structures, and repetitive sentence patterns, which support comprehension while allowing learners to focus on pronunciation and prosodic features during shadowing. The themes of the stories correspond to textbook units such as daily routines, interpersonal interaction, emotional expression, and community life, thereby ensuring continuity between curricular instruction and supplementary activities. From a pedagogical perspective, these books were also well suited for dramatization because they present clear narrative sequences, identifiable character roles, and dialogue-rich episodes. Such features facilitate role distribution, embodied language use, and interactive performance, which are essential for integrating dramatization with subsequent shadowing practice.

The books included the following: *The New Teacher* follows a robot's journey to become a competent primary school teacher, reinforcing expressions related to self-introduction and personal characteristics. *Zob Is Bored* demonstrates the daily routines of an alien and emphasizes the target structure "What do you do on weekends?" through cross-cultural dialogue. *The Street Party* depicts a neighborhood celebration and introduces vocabulary related to tableware, cooperation, and community. *Emma's Birthday* captures a child's emotional arc from disappointment to joy, supporting structures like "Are you going to...?" and expressions of surprise. *The Empty Room* features a mystery where the protagonist investigates strange events, promoting critical thinking and vocabulary for evidence and deduction. *Sophie is Really Angry* addresses emotional regulation and interpersonal communication, helping students express feelings and resolve conflict. *Toby and the Eagle* highlights environmental ethics and wildlife protection while teaching expressions related to nature and responsibility.

Procedure

During the 10-week period, both groups received the same in-school instruction based on the selected materials, along with an additional hour of English instruction per week. Before each new book, students received printed text for preview to familiarize themselves with the content. In class, the teacher introduced new language points and facilitated peer discussion.

For the experimental group, shadowing activities based on audio recordings of the picture books produced by professional native speakers were assigned as daily homework, rather than students' own dramatization recordings. In addition, all audio recordings were produced by the same native speaker and corresponded to the picture book texts used in class. Each shadowing task required students to complete the following steps:

1. Listening to native-speaker recordings of the story;
2. Shadowing the speech aloud at a matched pace;
3. Recording and submitting their oral performance via WeChat;
4. Receiving feedback from the teacher.

Students were expected to submit their recordings on a daily basis, and the teacher monitored participation and provided feedback on pronunciation and rhythm. Although recordings were used for instructional purposes, they were not archived for formal quantitative analysis.

The control group participated in all the same classroom activities but did not complete the shadowing assignments.

Data Analysis

The primary data for analysis consisted of students' scores from standardized English achievement assessments, including both overall English performance and listening comprehension subscores across the pre-test, post-test, and final-test. In addition to the primary focus on pre-test to post-test changes, final-test performance was examined as an exploratory indicator of retention. As all students completed the assessments at each measurement point, the dataset was complete across pre-test, post-test, and final-test, and no cases required exclusion due to missing data. Because the study employed a two-group, three-time-point design, the primary analytic approach was a mixed-design analysis of variance (mixed ANOVA) with group (experimental vs. control) as the between-subjects factor and time (pre-test, post-test, and final-test) as the within-subjects factor. This model was selected because it directly aligns with the study design and allows simultaneous examination of overall change over time, between-group differences, and most importantly, the group × time interaction, which indicates whether the two groups followed different performance trajectories across the intervention period. All

analyses were conducted using IBM SPSS Statistics 29.0. Descriptive statistics were first computed to summarize score distributions across groups and time points. Independent-samples t-tests were used to examine baseline equivalence between the two groups at pre-test. In addition, ANCOVA was conducted as a follow-up analysis to compare post-test outcomes while controlling for pre-test scores, thereby providing a more focused estimate of the intervention's effect on immediate learning. Paired-samples t-tests were used only as a supplementary within-group analysis to describe changes across measurement points. In this way, the mixed ANOVA served as the primary model for addressing the research questions, whereas the t-tests and ANCOVA were used to support the interpretation of specific contrasts.

Based on the research questions, the following hypotheses were formulated: Hypothesis 1: Chinese primary EFL learners who completed supplemental shadowing activities in conjunction with dramatization-enhanced instruction would demonstrate significantly higher overall English achievement than learners who received dramatization-based instruction only. Hypothesis 2: Learners in the experimental group (dramatization and shadowing) would show significantly greater improvements in listening comprehension scores than those in the control group (dramatization only).

RESULTS AND DISCUSSION

Descriptive Analysis

From Table 2, regarding English achievement, during the pre-test, the control group (CG) (M = 88.450, SD = 8.369) and the experimental group (EG) (M = 88.590, SD = 5.704) exhibited comparable mean scores, with highly overlapping distribution ranges (Control: 70–100; Experimental: 75–96). In the listening pre-test, the control group averaged 31.790 (SD = 6.737) while the experimental group averaged 30.280 (SD = 4.182). Although the experimental group showed slightly less variability, the difference in means was small. By the post-test, the experimental group's mean English score increased significantly to 93.210 (SD = 5.233), surpassing the control group's 88.660 (SD = 7.678). Concurrently, its minimum score increased to 79, indicating an overall upward shift in scores. In the listening post-test, however, the two groups' mean scores were similar (Control: M = 31.720, SD = 5.719; Experimental: M = 31.170, SD = 3.983), showing little difference. Final-test results revealed that in English performance, both groups maintained relatively high levels (Experimental: M = 89.900, SD = 7.512; Control: M = 88.90, SD = 5.984), but the difference between them narrowed. In the listening final-test, the control group averaged 31.410 (SD = 4.664), and the experimental group averaged 30.170 (SD = 3.694), indicating that the intergroup difference in listening scores remained limited. Overall, the experimental group showed higher post-test mean scores in overall English achievement, whereas listening gains remained modest.

TABLE 2
Descriptive Analysis

Measure	Group	Number	Minimum	Maximum	*Mean*	*SD*
English Pre-test	CG	29	70	100	88.450	8.369
	EG	29	75	96	88.590	5.704
English Listening Pre-test	CG	29	19	40	31.790	6.737
	EG	29	20	39	30.280	4.182
English Post-test	CG	29	70	99	88.660	7.678
	EG	29	79	100	93.210	5.233
English Listening Post-test	CG	29	18	39	31.720	5.719
	EG	29	21	40	31.170	3.983
English Final-test	CG	29	73	98	88.900	5.984
	EG	29	71	100	89.900	7.512
English Listening Final-test	CG	29	21	36	31.410	4.664
	EG	29	21	39	30.170	3.694

The Effectiveness of Short-term Intervention

To examine the differences in achievement between the experimental and control groups across the testing phases, this study employed independent samples t-tests to compare English and listening scores at the pre-test, post-test, and final-test stages, supplemented by effect size (*d*) to assess the practical significance of any differences (see Table 3). The independent samples t-tests revealed no statistically significant differences between the two groups in either English scores (t = -0.073, df = 56, p = .942, d = -.019) or listening scores (t = 1.030, df = 56, p = .307, d = .271) at the pre-test stage. This indicates that both groups were at a comparable baseline level in overall English achievement and listening performance prior to the experimental intervention. At the post-test stage, however, a statistically significant difference emerged in English scores (t = -2.638, df = 56, p = .011, d = -.693). The mean score of the experimental group was significantly higher than that of the control group, with an effect size approaching a medium level. This result indicates a statistically significant post-test difference, although the magnitude of raw score change should be interpreted cautiously within the 100 point scale. In contrast, the intergroup difference in post-test listening scores remained non-significant (t = 0.426, df = 56, p = .672, d = .112), implying that the intervention had limited effectiveness specifically in improving overall listening test performance. Final-test results showed no significant differences between the groups in either English scores (t = -0.561, df = 56, p = .577, d = -.147) or listening scores (t = 1.124, df = 56, p = .266, d = .295). The small effect sizes suggest that any observable differences in listening test scores were small and not sustained at the group level. Overall, the teaching intervention demonstrated a significant short-term boost to overall English achievement, but it showed limited effectiveness in improving listening as measured by the standardized listening test and showed limited evidence of long-term retention. It is essential to clarify that for all analyses, the sign of Cohen's d reflects the direction of the mean difference based on the subtraction order; therefore, effect magnitude is interpreted using $|d|$, while the sign is retained to indicate direction.

TABLE 3

Intergroup Comparison across Time Points

Measure	*t*	*df*	*p*	*d*
English Pre-test	-0.073	56	.942	-.019
English Listening Pre-test	1.030	56	.307	.271
English Post-test	-2.638	56	.011	-.693
English Listening Post-test	0.426	56	.672	.112
English Final-test	-0.561	56	.577	-.147
English Listening Final-test	1.124	56	.266	.295

Intra-group Comparison Across Time Points

Prior to the within-group longitudinal analyses, baseline comparability between the experimental and control groups was documented using descriptive statistics of pretest English and listening scores (Table 4). The two groups exhibited highly similar mean scores and variability at baseline, supporting their comparability prior to the experimental intervention.

TABLE 4

Baseline Equivalence of Pretest Scores between Groups

Measure (Pretest)	EG Mean (*SD*)	CG Mean (*SD*)	*t*	*p*	*d*
English	88.590 (5.70)	88.450 (8.37)	-0.073	.942	-.019
Listening	30.280 (4.18)	31.790 (6.74)	1.030	.307	.271

To evaluate changes over time within each group, paired-samples t-tests were conducted for listening and overall English scores across the three time points (pre-test, post-test, and final-test; see Table 5). This analysis assessed whether achievement changed significantly between time points and reported Cohen's d to quantify practical significance. For the control group's listening scores, none of the pairwise comparisons were statistically significant. Specifically, the mean difference between the pre-test and post-test was 0.069 (SD = 4.644), t(28) = 0.080, p = .937, d = .015; between the pre-test and final-test was 0.379 (SD = 5.328), t(28) = 0.383, p = .704, d = 0.071; and between the post-test and final-test was 0.310

(*SD* = 4.614), *t*(28) = 0.362, *p* = .720, *d* = .067. Overall, effect sizes were close to zero, suggesting that the control group's listening performance remained stable throughout the study.

TABLE 5
Paired-Samples T-test for Listening across Time in the Control Group

Comparison	*MD*	*SD*	*t*	*df*	*p*	*d*
Listening Pre-test vs. Post-test	0.069	4.644	0.080	28	.937	.015
Listening Pre-test vs. Final-test	0.379	5.328	0.383	28	.704	.071
Listening Post-test vs. Final-test	0.310	4.614	0.362	28	.720	.067

For the English scores of the control group, paired samples t-tests indicated that differences across time points were similarly non-significant (see Table 6). Specifically, the mean difference between pre-test and post-test was -0.207 (*SD* = 1.840, *t*(28) = -0.606, *p* = .550, *d* = -.112), suggesting negligible change. The mean difference between the pre-test and final-test was -0.448 (*SD* = 5.200, *t*(28) = -0.464, *p* = .646, *d* = -.086), representing a minimal difference. The mean difference between the post-test and final-test was -0.241 (*SD* = 4.673, *t*(28) = -0.278, *p* = .783, *d* = -.052), which was also statistically non-significant. The effect sizes (d) for all three comparisons were close to zero, indicating that the control group's English achievement remained essentially stable throughout the experiment, with no clear systematic upward or downward trend.

TABLE 6
Paired-Sample T-test Result for English Scores across Time in the Control Group

Comparison	*MD*	*SD*	*t*	*df*	*p*	*d*
English Pre-test vs. Post-test	-0.207	1.840	-0.606	28	.550	-.112
English Pre-test vs. Final-test	-0.448	5.200	-0.464	28	.646	-.086
English Post-test vs. Final-test	-0.241	4.673	-0.278	28	.783	-.052

For the listening scores of the experimental group, paired samples t-tests revealed statistically significant phase-specific changes across time points (see Table 7). The mean difference between pre-test and post-test was -0.897 (*SD* = 1.081, *t*(28) = -4.468, *p* < .001, *d* = -.830). Even though the pre-post change in listening was statistically significant, the absolute gain was modest, amounting to a 0.897-point increase on a 40-point test scale, approximately 2.2%. Therefore, this result should be interpreted as a small short-term increase in score rather than a substantial improvement in listening achievement. From an educational perspective, such a magnitude should be interpreted as a limited immediate improvement rather than a substantial enhancement of listening performance. The relatively large standardized effect (Cohen's *d*) should be interpreted cautiously, as a large *d* can occur when within-group variability (or the variability of difference scores) is small, even if the raw mean change is limited. In this case, the effect size reflects a consistent short-term shift within the experimental group rather than a large instructional impact. The mean difference between the pre-test and final-test was 0.103 (*SD* = 1.372, *t*(28) = 0.406, *p* = .688, *d* = .075), showing no significant difference, which suggests that the initial listening gains were not sustained over time. Conversely, the post-final decline indicates that the short-term listening gain was not maintained at the final assessment, suggesting limited retention under the current implementation conditions. The mean difference between post-test and final-test was 1.000 (*SD* = 0.845, *t*(28) = 6.372, *p* < .001, *d* = 1.183), indicating a statistically significant decline following the intervention period. Collectively, the experimental group exhibited significant short-term improvement in listening performance immediately following the intervention, but showed limited retention, with observable regression over time.

TABLE 7
Paired-Sample T-test Results for Listening across Time in Experiment Group

Comparison	*MD*	*SD*	*t*	*df*	*p*	*d*
Listening Pre-test vs. Post-test	-0.897	1.081	-4.468	28	< .001	-.830
Listening Pre-test vs. Final-test	0.103	1.372	0.406	28	.688	.075
Listening Post-test vs. Final-test	1.000	0.845	6.372	28	< .001	1.183

Performance Decay From Post-test to Final-test in the Experimental Group

For the English scores of the experimental group, paired samples t-tests demonstrated statistically significant phase-specific changes across time points (see Table 8). The mean difference between pre-test and post-test was -4.621 (SD = 5.906, t(28) = -4.213, p < .001, d = -.782). This statistically significant difference indicates that post-test scores were higher than pre-test levels, reflecting a statistically significant short-term increase in English test achievement within the experimental group. Importantly, a statistically significant performance decay was observed following the intervention period. Conversely, the mean difference between the post-test and final-test was 3.310 (SD = 5.621, t(28) = 3.171, p = .004, d = .589). This significant difference with a medium effect size reflects that final-test scores were significantly lower than post-test scores, indicating measurable attenuation of gains. Collectively, while the experimental group achieved a significant improvement in English performance during the intervention phase, partial regression occurred subsequently, suggesting that long-term retention was less pronounced than short-term gains. However, the mean difference between pre-test and final-test was -1.310 (SD = 7.106, t(28) = -0.993, p = .329, d = -.184), showing no statistical significance. This suggests limited sustainability of the intervention effects over an extended duration.

TABLE 8

Paired-Samples T-test Result for English Score across Time in the Experiment Group

Comparison	*MD*	*SD*	*t*	*df*	*p*	*d*
English Pre-test vs. Post-test	-4.621	5.906	-4.213	28	< .001	-.782
English Pre-test vs. Final-test	-1.310	7.106	-0.993	28	.329	-.184
English Post-test vs. Final-test	3.310	5.621	3.171	28	.004	.589

To examine short-term learning gains, ANCOVAs were conducted using post-test scores as the outcome variables, group (experimental vs. control) as the between-subject factors, and corresponding pretest scores as covariates. Full ANCOVA results, including Group, Pretest, and Group × Pretest effects, are reported in Table 9. For English post-test performance, a significant Group × Pretest interaction was observed, F(1, 54) = 12.48, p < .001, ηp^2 = .19, indicating that group differences varied as a function of initial English achievement. Building on these results, conditional analyses (Table 10) showed that the experimental group demonstrated larger advantages over the control group at lower and average pretest levels, whereas this advantage diminished at higher pretest levels. The adjusted main effect of group was also significant, F(1, 54) = 14.96, p < .001, ηp^2 = .22. In contrast, for listening post-test achievement, the Group × Pretest interaction did not reach statistical significance, F(1, 54) = 3.73, p = .058, ηp^2 = .07, and the adjusted group effect was not significant, F(1, 54) = 3.23, p = .078, ηp^2 = .06, suggesting that short-term group differences in listening were limited.

TABLE 9

ANCOVA Results for English and Listening Outcomes with Group × Pretest Interaction

Outcome	Source	*SS*	*df*	*MS*	*F*	*p*	ηp^2	*df* (error)
English Post-test	Group	195.547	1	195.547	14.962	<.001	.217	54
	Pre-test	1576.741	1	1576.741	120.646	<.001	.691	54
	Group × Pretest	163.032	1	163.032	12.475	<.001	.188	54
	Residual	705.735	54	13.069				54
Listening Post-test	Group	27.081	1	27.081	3.231	.078	.056	54
	Pre-test	492.797	1	492.797	58.802	<.001	.521	54
	Group × Pretest	31.245	1	31.245	3.728	.058	.065	54
	Residual	452.552	54	8.381				54
English Final-test	Group	0.209	1	0.209	0.007	.934	.000	54
	Pre-test	620.736	1	620.736	20.395	<.001	.274	54
	Group × Pretest	0.515	1	0.515	0.017	.897	.000	54
	Residual	1643.538	54	30.436				54
Listening Final-test	Group	61.669	1	61.669	7.986	.007	.129	54
	Pre-test	231.584	1	231.584	29.988	<.001	.357	54
	Group × Pretest	59.26	1	59.26	7.674	.008	.124	54
	Residual	417.017	54	7.723				54

To evaluate whether these effects were sustained over time, parallel ANCOVAs were conducted using final-test scores while controlling for pretest achievement, with the complete ANCOVA results again summarized in Table 9. For English final-test scores, neither the Group × Pretest interaction nor the adjusted group effect was significant (both $p > .89$), suggesting that the short-term English advantage did not persist at the final assessment. For listening final-test achievement, however, a significant Group × Pretest interaction was found, $F(1, 54) = 7.67$, $p = .008$, $\eta p^2 = .12$, and the main effect of group was also significant, $F(1, 54) = 7.99$, $p = .007$, $\eta p^2 = .13$. (see Table 9), indicating that intervention effects depended on students' initial listening achievement. Consistent with this pattern, conditional effects reported in Table 10 revealed that both the direction and magnitude of group differences varied across pretest scores. The experimental group performed slightly worse than the control group, whereas at higher pretest scores, the experimental group showed a small advantage. To support the validity of these findings, additional diagnostic checks and robustness analyses using heteroskedasticity-consistent HC3 standard errors are reported in Appendix Tables S1 and S2 (see Appendix).

TABLE 10

Conditional Group Differences at Selected Pretest Levels

Outcome	Pretest value	Mean (CG)	Mean (EG)	EG − CG
English Post-test	81.418	82.352	90.449	8.098
	88.517	88.717	93.18	4.463
	95.616	95.082	95.911	0.829
Listening Final-test	25.424	28.695	26.115	−2.580
	31.034	31.09	30.807	−0.283
	36.645	33.485	35.499	2.014

Note. Pretest values represent the mean and ±1 SD of the pretest distribution. Values are predicted means from the interaction model.

Total English Scores

As reported in Table 11, total English scores showed a significant Time × Group interaction, $F(2, 112) = 5.627$, $p = .005$, $\eta p^2 = .091$, indicating that EG and CG changed differently over time. There was also a main effect of Time, $F(2, 112) = 6.138$, $p = .003$, $\eta p^2 = .099$.

Follow-up comparisons with Bonferroni adjustment (Table 12) showed that EG improved from pre-test to post-test, $t(28) = -4.21$, $p = .001$, $|d| = .84$, but declined from post-test to final-test, $t(28) = 3.17$, $p = .011$, $d = .51$. CG showed no significant within-group changes over time (all adjusted $p = 1.00$). Between-group comparisons indicated that EG scored higher than CG at post-test, $t(56) = -2.64$, $p = .032$, $d = .69$, whereas the groups did not differ at pre-test or final-test (both adjusted $p = 1.00$). Overall, the intervention was associated with a short-term improvement in English performance that falls within the range of gains typically observed over a limited instructional period, and was not fully maintained one month after the intervention ended.

TABLE 11

Mixed-Design ANOVA (3 Time × 2 Group) for Total English Scores

Source	SS	*df1*	*df2*	MS	*F*	*p*	ηp^2	ε
Group	156.466	1	56	156.466	1.392	.243	.024	—
Time	173.115	2	112	86.557	6.138	.003	.099	0.929
Time × Group	158.724	2	112	79.362	5.627	.005	.091	—

TABLE 12

Simple Effects Follow-up for the Significant Time × Group Interaction (English)

Comparison	*t(df)*	*p(Bonf.)*	*Cohen's d*
EG: Pre vs Post	−4.21 (28)	.001	.84
EG: Post vs Final	3.17 (28)	.011	.51
CG: all within-group comparisons	—	ns (all = 1.00)	—
Post: CG vs EG	−2.64 (56)	.032	.69
Pre & Final: CG vs EG	—	ns (both = 1.00)	—

Listening Scores

Listening scores did not show comparable effects (Table 13). Neither the main effect of Time, $F(2, 112) = 1.019$, $p = .364$, $\eta p^2 = .018$, nor the Time × Group interaction, $F(2, 112) = 0.574$, $p = .565$, $\eta p^2 = .010$, was significant, and the main effect of Group was also not significant, $F(1, 56) = 0.869$, $p = .355$, $\eta p^2 = .015$. These results suggest that listening performance remained broadly similar across time in both groups. Although paired comparisons revealed significant short-term changes within the experimental group, the mixed-design ANOVA did not yield a significant main effect of Time or a Time × Group interaction, which may reflect a pattern in which the initial post-test gain was followed by regression at the final-test, thereby attenuating the overall three-time-point effect.

TABLE 13
Mixed-Design ANOVA (3 Time × 2 Group) for Listening Scores

Source	SS	*df1*	*df2*	MS	*F*	*p*	ηp^2	ε
Group	52.966	1	56	52.966	0.869	.355	.015	—
Time	12.736	2	112	6.368	1.019	.364	.018	0.961
Time × Group	7.172	2	112	3.586	0.574	.565	.010	—

This study examined the impact of two instructional conditions on students' overall English achievement and listening scores through a comparative analysis of control and experimental groups. Compared to the control group, the experimental group exhibited no significant advantage in either overall English or listening performance at the final-test stage. The mixed-design analyses indicated an outcome-specific intervention pattern: English scores showed a significant Time × Group interaction across the three measurement points, whereas listening scores did not show a significant Time × Group interaction. Overall, the mixed-design analyses suggest that the observed gains in English performance should be interpreted as modest short-term test-score improvements within the 100-point Tongkao scale, rather than as evidence of large or durable instructional effects. While the inclusion of a final-test allowed for the examination of retention over time, the findings related to retention should be interpreted with caution, given the observed attenuation of effects. The fading pattern observed at the final-test is consistent with evidence that learning gains from time-limited practice can attenuate when opportunities for spaced reactivation are reduced. In second language research, spaced practice has been shown to yield more robust delayed retention than massed practice, with a meta-analysis indicating a reliable advantage of spacing across L2 learning outcomes (Kim & Webb, 2023). More recent work further suggests that distributing practice over time can support the development and maintenance of L2 fluency when evaluated with delayed post-tests (Kakitanni & Kormos, 2024). Viewed through this perspective, the follow-up decline highlights the importance of sustained, curriculum-embedded re-engagement rather than suggesting that dramatization or shadowing is ineffective.

Pedagogically, brief but recurring shadowing cycles (e.g., weekly re-shadowing of key dialogue segments) aligned with textbook units may help maintain retrieval opportunities and reduce wash-out by extending practice beyond the intervention window. It is also important to consider that the listening measure used in the present study yielded only an overall listening score, rather than item-level subscores by question type. As a result, a finer-grained sub-analysis (e.g., distinguishing between literal and inferential comprehension) could not be conducted at the individual learner level. Importantly, although shadowing is theoretically expected to yield localized gains in lower-level auditory processing (e.g., segmentation, rhythm sensitivity, or perception-production alignment), the present study was unable to empirically examine such localized effects because the listening assessment did not provide item-level scores or differentiated subskill categories (e.g., literal vs. inferential comprehension). In addition, the listening items were designed to assess curriculum-aligned comprehension under regular testing conditions and were thus relatively homogeneous in format and focus. As a result, such localized gains may not be readily observable when listening performance is operationalized as a single composite score, which may partly account for the absence of statistically significant group differences in listening outcomes.

In terms of RQ2, the absence of statistically significant improvements in listening comprehension does not necessarily imply that shadowing is not theoretically useful. From a sociocultural perspective, the instructional sequence in the present study may be understood as a scaffolded progression from socially supported language use to increasingly self-regulated performance within learners' Zone of Proximal Development (ZPD). During dramatization activities, learners engaged in collaborative interaction in which meaning, prosody, and discourse patterns were jointly constructed under teacher and peer guidance. Such contexts provide social mediation that enables learners to perform beyond their independent level. The subsequent transition to individual shadowing represents a gradual shift toward self-regulation, as learners rehearse socially

encountered language in a more private, self-directed manner. Through repeated rehearsal, learners may consolidate linguistic forms previously encountered in socially mediated contexts, contributing to the consolidation of prosodic control and speech processing routines that later support independent comprehension. This interpretation is supported by prior EFL research suggesting that shadowing promotes active engagement with auditory input rather than passive reception. For example, Jeon (2011) found that learners who engaged in shadow reading demonstrated significantly greater gains in listening comprehension, which were attributed to the simultaneous processing and reproduction of speech, including rhythm, stress, and intonation. This interpretation frames dramatization and shadowing as a sequenced instructional design in which socially mediated classroom interaction is followed by individual rehearsal that may support greater self-regulation over time.

One possible explanation for the non-significant listening outcomes concerns the developmental gap between phonological processing and its application to general listening comprehension, particularly among young EFL learners. Although shadowing is believed to enhance prosodic awareness and strengthen the perception-production link, such low-level auditory refinements may require sustained exposure before becoming observable in broad listening assessments. A second consideration relates to the nature of the listening measure employed in the present study. The Tongkao listening test is designed to assess general curriculum-oriented comprehension rather than fine-grained prosodic sensitivity or fluency-related processing skills. Consequently, subtle improvements in rhythm, stress perception, or speech processing efficiency—areas most directly targeted by shadowing may not be fully captured by this assessment.

Another contextual factor that warrants consideration is learners' extracurricular English exposure beyond the classroom and assigned shadowing activities. Although the two groups were drawn from the same intact class and shared the same school environment, it is possible that individual differences in access to private tutoring, family support, or additional English resources outside school may have contributed to variability in learning outcomes. Such influences are common in school-based research and are often difficult to isolate within regular instructional settings. The present study did not include a formal survey of students' out-of-school English learning experiences; therefore, these background factors could not be statistically controlled. However, because participants were drawn from the same class and instructional context, any extracurricular exposure cannot be assumed to have systematically favored one group. These considerations suggest that the findings should be interpreted within the ecological conditions of regular schooling rather than as outcomes produced under fully controlled experimental circumstances.

A further consideration relates to the nature of practice implementation in the present study. Although task completion was monitored through audio submission, these recordings were used for instructional purposes and were not archived or coded as analyzable research data, and thus could not be retrospectively analyzed. Because the shadowing activities were completed as home-based assignments, detailed information regarding individual practice frequency, duration, and performance accuracy could not be systematically quantified. Similar constraints are frequently acknowledged in studies involving out-of-class or mobile-assisted language practice, where instructional activities occur under more naturalistic conditions rather than tightly controlled experimental settings. In such contexts, variability in learner engagement forms part of the instructional ecology. Moreover, development in phonological and listening-related skills does not necessarily progress in a linear relation to practice quantity; gains may depend on the quality of attention during practice, individual learner differences, and possible threshold or plateau effects. From this perspective, the absence of fine-grained dosage indices complicates precise causal attribution but does not in itself undermine the theoretical plausibility of shadowing-related learning processes. Instead, the findings may be understood as reflecting instructional outcomes under authentic implementation conditions and should be interpreted with appropriate caution. Therefore, the present study cannot fully determine the extent to which observed post-test gains were attributable specifically to the shadowing activity rather than other contextual influences.

CONCLUSION AND IMPLICATION

The present study explored the potential effects of combining in-class dramatization with out-of-class shadowing activities on the English language achievement of Chinese primary school learners. Findings indicated that while the experimental group, which received daily shadowing assignments in addition to dramatization-based instruction, demonstrated modest short-term improvements in Tongkao total scores compared to the control group, no statistically significant gains were observed in listening comprehension. Moreover, the initial advantage in total English scores for the experimental group appeared to diminish by the final-test, indicating attenuation of the initial gains over time. These results offer a more nuanced perspective on prior research highlighting the potential of shadowing to enhance listening comprehension, particularly in

secondary and adult EFL contexts (Ekayati, 2020; Hamada, 2020; Zuhriyah, 2016). The findings suggest that integrating shadowing activities may have limited, context-dependent pedagogical relevance at the primary level, particularly when evaluated using curriculum-based achievement measures, as reflected in the short-term gains in English test scores observed in the present study. Shadowing assignments, as implemented in the present study, may function more effectively as a supplementary practice format rather than a standalone solution. Taken together, these findings indicate that RQ1 was partially supported in terms of short-term gains in English achievement, whereas RQ2 was not supported, as no significant improvements were observed in listening comprehension.

From a classroom practice perspective, shadowing appears to work best when it functions as a brief extension of interactive classroom activities rather than an isolated homework routine. When the shadowing materials are drawn directly from previously dramatized dialogues, learners encounter familiar expressions, narrative contexts, and prosodic patterns. This continuity may help younger learners maintain engagement while reducing processing demands. In this sense, short and regular shadowing practice linked to story-based dramatization can reinforce previously enacted language and extend exposure beyond the limited time available for classroom interaction.

The implications for future research and instructional design emerge from this study. Extending the duration of shadowing-based interventions beyond 10 weeks is recommended to examine whether long-term engagement produces more stable gains in both listening and other language domains. Future studies should adopt a more comprehensive assessment framework that includes additional language domains such as speaking, reading, and writing to better capture the full spectrum of shadowing's impact. Additional qualitative data, such as learner reflections or teacher observations, could further illuminate how students experienced the dramatization-shadowing sequence, particularly in relation to motivation and perceived task difficulty. Expanding the sample size and incorporating participants from multiple schools or regions could enhance the generalizability of the findings. Lastly, more detailed process-oriented research, such as analyzing learners' recordings to assess prosodic development or engagement levels, could help identify specific mechanisms by which shadowing contributes to language acquisition.

In sum, while shadowing alone may not be regarded as a universal solution to all language learning challenges at the primary level, it may still contribute to a more balanced instructional ecology when combined with contextualized classroom instruction, such as dramatization. The findings should therefore be interpreted as context-bound associations observed under regular schooling conditions rather than isolated causal effects. In particular, the absence of quantitative compliance measures prevents precise attribution of observed gains to the shadowing component itself.

LIMITATIONS

The findings of this study should be interpreted with several limitations. As discussed earlier, the relatively short duration of the intervention (10 weeks) may have been insufficient for younger learners to exhibit stable and measurable improvements in listening subskills, which often develop gradually over extended periods of exposure and repetition. The observed short-term gains followed by post-intervention regression suggest limited retention under the current implementation conditions. Another limitation relates to the study design. The participants were assigned to experimental and control conditions within a single intact class. Because the assignment was conducted at the within-class level rather than through random selection of intact classes or population-level sampling, background characteristics could not be fully balanced through randomization.

In addition, the study did not collect pre-assignment data on family-level background variables such as socioeconomic status, parental involvement, home English exposure, or device conditions. These contextual factors are common sources of variability in school-based research and could not be statistically controlled in the present design. Consequently, it is possible that some of the observed short-term differences reflect variation in home support or learning conditions rather than the instructional intervention alone. For example, if students with stronger parental involvement or more stable device access engaged more consistently with the home-based shadowing activities, the estimated intervention effect may be inflated. Conversely, uneven access or limited support may have attenuated potential gains by introducing additional variability in homework engagement. Because both groups were drawn from the same intact classroom and shared the same school environment, systematic between-group imbalances are less likely, but they cannot be ruled out without direct measurement. The findings should therefore be interpreted within the ecological conditions of regular schooling rather than as effects isolated from all contextual influence. With regard to the shadowing intervention, students were required to submit audio recordings via a WeChat study group, which were used to verify task completion and to provide instructional feedback. However, the recordings were not systematically coded as research data because they were collected primarily for

instructional monitoring rather than for research documentation. As a result, they were unavailable for retrospective quantitative coding of practice frequency, duration, or performance accuracy. This limitation constrained a more fine-grained examination of intervention dosage and potential dose-response relationships. Consequently, the study cannot determine whether higher levels of engagement would have been associated with larger or more durable learning gains.

Lastly, the study was conducted in a single school and classroom, which may limit the generalizability of the findings to other educational contexts. Future research employing longer interventions, multi-site samples, and more fine-grained process measures is needed to further validate and extend the present findings.

References

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial Short-Term and Working Memory in Children: Are they separable? *Child Development*, *77*(6), 1698–1716. https://doi.org/10.1111/j.1467-8624.2006.00968.x

Baddeley, A. (2006). Working memory: An overview. In S. J. Pickering (Ed.), *Working memory and education* (pp. 1–31). Elsevier. https://doi.org/10.1016/B978-012554465-8/50003-X

Butler, Y. G. (2015). English language education among young learners in East Asia: A review of current research (2004–2014). *Language Teaching*, *48*(3), 303–342. https://doi.org/10.1017/s0261444815000105

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, *25*(1), 15–37. https://doi.org/10.1177/0265532207083743

Chou, C. (2013). A study on the effectiveness of applying "readers' theater" as English remedial instruction for underachievers. *Taiwan Journal of TESOL*, *10*(1), 77–103.

Diamond, A. (1995). Evidence of robust recognition memory early in life even when assessed by reaching behavior. *Journal of Experimental Child Psychology*, *59*(3), 419–456. https://doi.org/10.1006/jecp.1995.1020

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dong, M., Fan, J., & Xu, J. (2023). Differential washback effects of a high-stakes test on students' English learning process: Evidence from a large-scale stratified survey in China. *Asia Pacific Journal of Education*, *43*(1), 252–269. https://doi.org/10.1080/02188791.2021.1918057

Ekayati, R. (2020). Shadowing technique on students' listening word recognition. *IJEMS: Indonesian Journal of Education and Mathematical Science*, *1*(2), 31–42. https://doi.org/10.30596/ijems.v1i2.4695

Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics*, *19*(3), 221–246. https://doi.org/10.1111/j.1473-4192.2009.00231.x

Feng, A. (2012). Spread of English across greater China. *Journal of Multilingual and Multicultural Development*, *33*(4), 363–377. https://doi.org/10.1080/01434632.2012.661435

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, *3*(1), 34–56. https://doi.org/10.1075/jslp.3.1.02foo

Gill, R., Barbour, J., & Dean, M. (2014). Shadowing in/as work: Ten recommendations for shadowing fieldwork practice. *Qualitative Research in Organizations and Management: An International Journal*, *9*(1), 69–89. https://doi.org/10.1108/QROM-09-2012-1100

Hamada, Y. (2014). The effectiveness of pre- and post-shadowing in improving listening comprehension skills. *The Language Teacher*, *38*(1), 3–8. https://doi.org/10.37546/jalttlt38.1-1

Hamada, Y. (2015). Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension. *Language Teaching Research*, *20*(1), 35–52. https://doi.org/10.1177/1362168815597504

Hamada, Y. (2018). Shadowing: what is it? How to use it. Where will it go? *RELC Journal*, *50*(3), 386–393. https://doi.org/10.1177/0033688218771380

Hamada, Y. (2020). Developing a new shadowing procedure for Japanese EFL learners. *RELC Journal*, Advance online publication. https://doi.org/10.1177/0033688220937628

He, A. E. (2011). Educational decentralization: A review of popular discourse on Chinese–English bilingual education. *Asia Pacific Journal of Education*, *31*(1), 91–105. https://doi.org/10.1080/02188791.2011.544245

Hu, G. (2005). English language education in China: Policies, progress, and problems. Language *Policy*, *4*(1), 5–24. https://doi.org/10.1007/s10993-004-6561-7

Irby, B. T., Sue, C. A., Smith, K. M., & Alwan, M. (2016). Maximizing the shadowing experience: A guidance document. *Hospital Pharmacy*, *51*(1), 54–59. https://doi.org/10.1310/hpj5101-54

Jeon, E. (2011). The effects of shadowing reading on EFL learners' English listening comprehension. *Modern English Education*, *12*(4), 277–296.

Jin, L., & Cortazzi, M. (2018). Early English language learning in East Asia. In L. Jin & M Cortazzi (Eds.), *Researching intercultural learning: Investigations in language and education* (pp. 477–492). Routledge.

Lambert, S. (1992). Shadowing. *Meta*, *37*(2), 263–273. https://doi.org/10.7202/003378ar

Kakitanni, J., & Kormos, J. (2024). The effects of distributed practice on second language fluency development. *Studies in Second Language Acquisition*, *46*, 770–794. https://doi.org/10.1017/S0272263124000251

Kim, S. K., & Webb, S. (2023). Does spaced practice have the same effects on different second language vocabulary learning activities? Fill-in-the-blanks versus flashcards. *The Modern Language Journal*, *107*(4), 944–964. https://doi.org/10.1111/modl.12879

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Liu, J. (2002). Process drama in second- and foreign-language classrooms. In G. Brauer (Ed.), *Body and language: Intercultural learning through drama* (pp. 1–25). Ablex.

Ma, S., & Liu, J. (2022). A Field study on the script structure of English drama in education in Chinese primary and secondary school. *English Language and Literature Studies*, *12*(4), 16. https://doi.org/10.5539/ells.v12n4p16

Nation, P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt039.0

Ozverir, I., Herrington, J., & Osam, U. V. (2016). Design principles for authentic learning of English as a foreign language. *British Journal of Educational Technology*, *47*(3), 484–493. https://doi.org/10.1111/bjet.12449

Parker, J., Hughes, M., & Rutter, L. (2006). "Shadowing" and its place in preparing students for practice learning. *The Journal of Practice Teaching and Learning*, *7*(3), 49–69. https://doi.org/10.1921/19643

Pereira, L. C., Vieira, F., & Teófilo, A. (2019). Expressive reading and dramatization of stories in teaching English to young learners. *CLELEjournal*, *7*(1), 45–60.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142–173. https://doi.org/10.1191/0265532205lt300oa

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education: Principles, Policy & Practice*, *14*(1), 51–74. https://doi.org/10.1080/09695940701272856

Saito, K. (2013). Experienced teachers' perspectives on priorities for improved intelligible pronunciation: The case of Japanese learners of English. *International Journal of Applied Linguistics*, *24*(2), 250–277. https://doi.org/10.1111/ijal.12026

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, *55*(3). https://doi.org/10.1002/tesq.3027

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Serrurier-Zucker, C., & Gobbé-Mévellec, E. (2014). The page is the stage: From picturebooks to drama with young learners. *CLELEjournal*, *2*(2), 13–30. https://doi.org/10.5281/zenodo.322587

Stinson, M., & Freebody, K. (2006). The dol project: The contributions of process drama to improved results in English oral communication. *Youth Theatre Journal*, *20*(1), 27–41. https://doi.org/10.1080/08929092.2006.10012585

Stinson, M., & Winston, J. (2011). Drama education and second language learning: a growing field of practice and research. *Research in Drama Education: The Journal of Applied Theatre and Performance*, *16*(4), 479–488. https://doi.org/10.1080/13569783.2011.616395

Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *Canadian Modern Language Review*, *50*(1), 158–164. https://doi.org/10.3138/cmlr.50.1.158

Swain, M., & Watanabe, Y. (2019). Languaging: Collaborative dialogue as a source of second language learning. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–8). Wiley. https://doi.org/10.1002/9781405198431.wbeal0664.pub2

Takeuchi, H., Maruyama, T., Taki, Y., Motoki, K., Jeong, H., Kotozaki, Y., Shinada, T., Nakagawa, S., Nouchi, R., Iizuka, K., Yokoyama, R., Yamamoto, Y., Hanawa, S., Araki, T., Sakaki, K., Sasaki, Y., Magistro, D., & Kawashima, R. (2021). Effects of training of shadowing and reading aloud of second language on working memory and neural systems. *Brain Imaging and Behavior*, 15, 1253–1269. https://doi.org/10.1007/s11682-020-00324-4

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3). https://doi.org/10.1111/lang.12343

Ulas, A. H. (2008). Effects of creative, educational drama activities on developing oral skills in primary school children. *American Journal of Applied Sciences*, *5*(7), 876–880. https://doi.org/10.3844/ajassp.2008.876.880

Winston, J., & Stinson, M. (2014). Drama education and second language learning. In J. Winston (Ed.), *Second language learning through drama: Practical techniques and applications* (pp. 1–10). Routledge.

Wu, H.-F. (2014). *Effects of process drama-assisted intervention on oral communication strategies* (Doctoral dissertation, Griffith University).

Yang, J. (2006). Learners and users of English in China. *English Today*, *22*(2), 3–10. https://doi.org/10.1017/s0266078406002021

Zuhriyah, M. (2016). Improving students' listening skill through shadowing. *Register Journal*, *9*(2), 124–136. https://doi.org/10.18326/rgt.v9i2.703

Appendix

Assumption Checks and Robust Estimates for ANCOVA Models

Table S1 and Table S2 are included to document model diagnostics and to demonstrate the robustness of the ANCOVA results. Table S1 reports assumption checks for each ANCOVA model, including tests of residual normality and homogeneity of variance. As shown in Table S1, several models, particularly those involving listening outcomes and final-test scores, exhibited evidence of heteroskedasticity. Such patterns are common in educational data and therefore motivate the use of additional robustness checks in statistical inference. To address this issue, Table S2 presents coefficient estimates from the ANCOVA interaction models using heteroskedasticity-consistent HC3 standard errors. These robust estimates reduce sensitivity to unequal variances and allow readers to evaluate whether the main conclusions remain unchanged under more conservative assumptions. The coefficients reflect adjusted group differences between the experimental and control groups, while interaction terms indicate whether group effects varied as a function of students' initial achievement.

TABLE S1

Assumption Checks for ANCOVA Models

Outcome (DV)	Covariate	*N*	R^2	Adj. R^2	Shapiro–*W*	Shapiro *p*	BP *F p*
English Post-test	english_pre	58	0.74	0.726	0.890	<0.001	0.089
Listening Post-test	listening_pre	58	0.668	0.65	0.941	0.007	0.001
English Final-test	english_pre	58	0.367	0.332	0.971	0.177	0.006
Listening Final-test	listening_pre	58	0.589	0.566	0.953	0.025	<0.001

TABLE S2

Robust (HC3) Coefficient Estimates for ANCOVA Interaction Models

Outcome (DV)	Term	Coef.	SE (HC3)	*p*	95% CI
English Post-test	Group (EG)	49.779	21.751	.026	[6.171, 93.386]
	Pretest	0.897	0.081	<.001	[0.734, 1.059]
	Group × Pretest	−0.512	0.244	.041	[−1.002, −0.022]
Listening Post-test	Group (EG)	−8.608	3.967	.034	[−16.561, −0.654]
	Pretest	0.623	0.110	<.001	[0.402, 0.844]
	Group × Pretest	0.297	0.120	.016	[0.057, 0.538]
English Final-test	Group (EG)	−1.627	27.180	.952	[−56.118, 52.865]
	Pretest	0.563	0.174	.002	[0.215, 0.911]
	Group × Pretest	0.029	0.299	.924	[−0.571, 0.629]
Listening Final-test	Group (EG)	−12.989	4.933	.011	[−22.879, −3.099]
	Pretest	0.427	0.124	<.001	[0.178, 0.676]
	Group × Pretest	0.409	0.148	.008	[0.112, 0.707]

Note. Group (EG) represents the estimated difference between the experimental group and the control group.